

Exercise 1: Kernelized Multiclass SVM

For the data set $\mathcal{D} = ((\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(n)}, y^{(n)}))$ with $y^{(i)} \in \mathcal{Y} = \{-1, 1\}$, assume we are provided with a suitable feature map $\phi : \mathcal{X} \rightarrow \Phi$, where $\Phi \subset \mathbb{R}^d$. In the featurized SVM learning problem we are facing the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \theta_0, \zeta^{(i)}} \quad & \frac{1}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + C \sum_{i=1}^n \zeta^{(i)} \\ \text{s.t.} \quad & y^{(i)} \left(\langle \boldsymbol{\theta}, \phi(\mathbf{x}^{(i)}) \rangle + \theta_0 \right) \geq 1 - \zeta^{(i)} \quad \forall i \in \{1, \dots, n\}, \\ \text{and} \quad & \zeta^{(i)} \geq 0 \quad \forall i \in \{1, \dots, n\}, \end{aligned}$$

where $C \geq 0$ is some constant.

- (a) Argue that this is equivalent to the following ERM problem:

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n \max(1 - y^{(i)} (\boldsymbol{\theta}^\top \phi(\mathbf{x}^{(i)}) + \theta_0), 0),$$

i.e., the regularized ERM problem for the hinge loss for the hypothesis space

$$\mathcal{H} = \{f : \Phi \rightarrow \mathbb{R} \mid f(\mathbf{z}) = \boldsymbol{\theta}^\top \mathbf{z} + \theta_0 \quad \boldsymbol{\theta} \in \mathbb{R}^d, \theta_0 \in \mathbb{R}\}.$$

- (b) Now assume we deal with a multiclass classification problem with a data set $\mathcal{D} = ((\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(n)}, y^{(n)}))$ such that $y^{(i)} \in \mathcal{Y} = \{1, \dots, g\}$ for each $i \in \{1, \dots, n\}$. In this case, we can derive a similar regularized ERM problem by using the multiclass hinge loss (see Exercise Sheet 4 (b)):

$$\mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{\theta}\|^2 + C \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y) - \boldsymbol{\theta}^\top \psi(\mathbf{x}^{(i)}, y^{(i)}), 0),$$

where $\psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^d$ is suitable (multiclass) feature map. Specify a ψ such that this regularized multiclass ERM problem coincides with the regularized binary ERM problem in (a).

- (c) Show that the regularized multiclass ERM problem in (b) can be written in the following kernelized form:

$$\frac{1}{2} \boldsymbol{\beta}^\top \mathbf{K} \boldsymbol{\beta} + C \sum_{i=1}^n \sum_{y \neq y^{(i)}} \max(1 + (\mathbf{K} \boldsymbol{\beta})_{(i-1)g+y} - (\mathbf{K} \boldsymbol{\beta})_{(i-1)g+y^{(i)}} , 0),$$

where $\boldsymbol{\beta} \in \mathbb{R}^{ng}$ and $\mathbf{K} = \mathbf{X} \mathbf{X}^\top$ for $\mathbf{X} \in \mathbb{R}^{ng \times d}$ with row entries $\psi(\mathbf{x}^{(i)}, y)^\top$ for $i = 1, \dots, n, y = 1, \dots, g$, i.e.,

$$\mathbf{X} = \begin{pmatrix} \psi(\mathbf{x}^{(1)}, 1)^\top \\ \psi(\mathbf{x}^{(1)}, 2)^\top \\ \vdots \\ \psi(\mathbf{x}^{(1)}, g)^\top \\ \psi(\mathbf{x}^{(2)}, 1)^\top \\ \vdots \\ \psi(\mathbf{x}^{(n)}, g)^\top \end{pmatrix}.$$

Here, $(\mathbf{K} \boldsymbol{\beta})_{(i-1)g+y}$ denotes the $((i-1)g+y)$ -th entry of the vector $\mathbf{K} \boldsymbol{\beta}$.

Hint: The representer theorem tells us that for the solution $\boldsymbol{\theta}^*$ (if it exists) of $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$ it holds that $\boldsymbol{\theta}^* \in \text{span}\{\psi(\mathbf{x}^{(i)}, y)_{i=1, \dots, n, y=1, \dots, g}\}$.

Exercise 2: Kernel Trick

The polynomial kernel is defined as

$$k(x, \tilde{x}) = (x^T \tilde{x} + b)^d.$$

Furthermore, assume $x \in \mathbb{R}^2$ and $d = 2$.

- (a) Derive the explicit feature map ϕ taking into account that the following equation holds:

$$k(x, \tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle$$

- (b) Describe the main differences between the kernel method and the explicit feature map.